PB-227 766

STEP BACKWARD REGRESSION

Gary I. Selner, et al

Federal Water Pollution Control
Administration
Annapolis, Maryland

August 1969

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

PB 227 766

Chesapeake Technical Support Laboratory
Middle Atlantic Region
Federal Water Pollution Control Administration
U.S. Department of the Interior

Technical Report No. 29

STEP BACKWARD REGRESSION

by

Gary I. Selner*

August 1969

*Revised by Paul R. Dorn

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

Johan A. Aalto, Chief, CTSL
Norbert A. Jaworski, Chief, Engineering Section
Richard Burkett, Draftsman

TABLE OF CONTENTS

	Page
Program	1
Abstract	1
Description	2
Restrictions	5
Program Run Preparation	6
JOB card	7
System Control Cards	10
Header Card	13
Options	14
Column Identifier Cards	15
Row Selector Cards	16
Column Selector Cards	17
Minimum Number of Variables Card	18
Residual Start Card	19
Output	20
Sample Output	21

Preceding page blank

PROGRAM: STEP BACKWARD REGRESSION

ABSTRACT: This program performs a multiple regression analysis and provides related statistics on a STATPAC data matrix or on a subset of that matrix formed by the selection of rows and columns of the matrix. If desired, the regression may proceed stepwise where, at each step, the least significant independent variable is deleted from the regression equation. The regression residuals may also be computed.

RESTRICTIONS:

The input matrix may have a maximum size of 99999 rows and 199 columns. The maximum size of the subset selected for a regression analysis is 99999 rows and 99 columns.

PROGRAM RUN PREPARATION:

The following is a complete deck setup for this program:

1. JOB card
2. System control cards
3. Header card
4. Column selector cards*
5. Row selector cards*
6. Minimum number of variables card *
7. Residual start card*
8. Delimeter

*Optional--see the individual sections on the various cards.

If more than one run is to be performed, repeat cards 3-7 as necessary; the final card should be delimiter card.

JOB CARD:

The JOB card cannot be catalogued and is installation as well as machine dependent. Below is the specification for the 360/Model 65 located in the Department of the Interior building, in Washington, D.C. The structure of the JOB card changes occasionally, and one should check with the computer center before using the below form.

Card 1 -

Card Columns	Contents
1 - 2	//
3	Center Code: D = Denver, Colorado F = Flagstaff, Arizona I = Crystal Plaza, Virginia M = Menlo Park, California R = Rolls, Missouri W = Washington, D.C.
4 - 5	Agency Code
6 - 8	User registration code.
9 - 10	User's ID. May be changed by the user as desired. Do not use the same two characters for different jobs run during the same day.
11	Blank.
12 - 14	JOB (i.e. the word JOB).

Card Columns	Contents
15	Blank.
16	((Left parenthesis).
17 - 20	Program Number
21	, (Comma).
22 - 25	Auxiliary account number.
26	, (Comma).
27 - 30	Estimated execution time in minutes. Requires four numeric digits.
31	, (Comma).
32 - 35	Estimated lines of print expressed in thousands of lines. Requires four numeric digits.
36	, (Comma).
37 - 40	Estimated number of cards to be punched. Requires four numeric digits.
41	, (Comma).
42	Reserved for future use; must be a 1 punch.
43	, (Comma).
44	Reserved for future use; must be a 1 punch.
45	, (Comma).
46	Type of Run: C = Compile only T = Test of program P = Production use of program

Card Columns	Contents
47	, (Comma).
48 - 49	Number of lines per page. A value of zero suppresses page overflow tests. If this field (and the preceding comma) is eliminated, a default option of 61 lines per page is used. In this case the following fields are shifted left 3 columns. (Except columns 62-72.)
50	) (Right parenthesis).
51	, (Comma).
52	' (Apostrophe).
53 - 61	Name of the user.
62	' (Apostrophe).
63	, (Comma).
64 - 71	Blank.
72	X (The letter X).

Card 2 -

Card Columns	Contents
1 - 2	//
3 - 15	Blank.
16 - 25	MSGLEVEL=1
26	, (Comma).
27 - 33	CLASS=C
34 - 72	Blank.

SYSTEM CONTROL CARDS:

Cards 1 & 2:

(a) If the object deck is used -

```
1
//bEXECbLINKFORT,REGION.GO=252K,TIME.GO=J
//LKED.SYSINbDDb*
```

where J is the time required to run the program(in minutes). The 'b' stands for a blank space. Next comes the object deck (including its delimiter).

(b) If the source deck is used -

```
1
//bEXECbFORTGCLG,PARM.FORT='DECK',REGION.GO=252K,TIME.GO=J
//FORT.SYSINbDDb*
```

where J is as before. Next comes the source deck (including its delimeter).

Cards 3 & 4:

(a) If the data resides on the disc SYSDK -

```
1
//GO.FT10F001bDDbDSN=&NAME,UNIT=SYSDY,DISP=(OLD,PASS/DELETE,DELETE),
//bDCB=(RECFM=VB,LRECL=RRR,BLKSIZE=BBBB)
```

where &NAME is the name of the storage space. (The '&' signifies the storage is temporary--it only exists for the extent of the job.) PASS is used if the data file is used later on; otherwise use DELETE.

The letters 'RRR' and 'BBBB' are computed as follows:

RRR = 8M + 24 where M = number of columns in the data matrix.

BBBB = K(RRR) + 4 where K is an integer chosen so that the positive difference (7200-BBBB) is as small as possible.

(b) If the data resides on magnetic tape -

```
1
//GO.FT10F001bDDbUNIT=2400,LABEL=(,SL),VOLUME=(,RETAIN,,,SER=YYYYYY),
//bDCB=(RECFM=VB,LRECL=RRR,BLDSIZE=BBBB),
//bDISP=(OLD,KEEP),DSP=STAPAC
```

where 'YYYYYY' in the first tape card represents a six digit input tape number (leading zeros must be given).

The letters 'RRR' and 'BBBB' are as before.

When a tape is used, a tape setup card is required. Its form is:

Card Columns	Contents
1 - 9	/*MESSAGE
10 - 12	Blank.
13 - 20	The same characters as in columns 3-10 of the JOB card.
21 - 22	Blank.
23 - 27	SETUP.
28 - 29	Blank.
30 - 36	The number of the tape used.
37 - 38	/9

If the tape is written on as well as read, in column 39, place an R (for ring in). This card should be placed right after the JOB card. (This card's format is particular to the 360/65 in Washington, D.C.)

Card 5:

1
//GO.SYSINbDDb*

HEADER CARD:

Columns	Format	Entry	Description
1-30	7A4,A2	TITLE	Up to 30 characters of alphanumeric information used to title the output for this data set. It is also used when listing the total number of plots created. It is not used on the graph.
31-38	2A4	INPUT ID	Up to 8 characters of alphanumeric information used to identify the input data set.
39-43	I5	INPUT N	The number of rows in the input data matrix. (Right justified.)
44-46	I3	INPUT M	The number of columns in the input data matrix. (Right justified, 199.)
47-56	10I1	OPTION	See the following sheet.
73-77	I5	PRON	The number of pairs of row numbers needed to select the desired rows of the input matrix. If blank, all rows are included. If not blank, this number must be right justified and row selector cards must be included.
78-80	I3	PROM	The number of pairs of column numbers needed to select the desired columns in the input data matrix. (If blank, all columns are included. If not blank, this number must be right justified and column selector cards must be provided.)

OPTIONS -

OPTION(1) 0 - No stepwise deletion of variables
-1 - Deletion of least significant variable occurs until the program reaches minimum specified.

OPTION(2) 0 - No action taken.
-1 - Residual analysis performed starting with all variables present; continues until stepwise deletion stops.

OPTION(3) -2 - Residual analysis performed starting with number of variables specified on Residual start card.

OPTION(4) 0 - Indeterminants not allowed in the selected data
-1 - Observations having indeterminants in selected data are skipped.

OPTION(5) 0 - No action taken
-1 - Residual analysis doen using antilog transform of dependent variable.

OPTION(6) - NOT USED

OPTION(7) - NOT USED

OPTION(8) - NOT USED

OPTION(9) 0 - Input tape unit is 10.
-1-9 This is the input tape unit.

OPTION(10) 0 - Use the variable identifiers on the STATPAC tape
-1 - Read in new variable identifiers from cards.

COLUMN IDENTIFIER CARDS:

These cards are used only if OPTION(10)=1 on the header card. They permit the user to associate an eight character identifier with each column in the data matrix. Ten identifiers can be used per card; if more than ten identifiers are needed, they are continued on another card. The numbers above the eight card column fields indicate the columns in the data matrix the identifiers are associated with. One identifier must be specified for each column in the output data matrix.

ROW SELECTOR CARDS:

These cards are used only if PRON on the header card is not blank. The number entered in the field PRON specifies how many pairs of row numbers are used to select desired rows of a data matrix. Each pair specifies the rows FROM and including the first member of the pair TO and including the last member of the pair will be selected. The pairs must be entered starting in the left most field of the card and continuing across eight pairs per card If more pairs are used, continue on another card. The first member of the pair is entered under the word "FROM" and the second member of the pair entered under "TO." Except for the selection of a single row (where the "TO" portion is left blank) the row numbers must form an increasing sequence, i.e., rows must be selected in the order they appear in the data matrix.

COLUMN SELECTOR CARDS:

These cards are used only if PRON on the header card is not blank. The number entered in the field PRON specifies how many pairs of column numbers are used to select desired columns of a data matrix. The instructions for their use is identical to the select row cards except that thirteen pairs per card are used and the columns need not be selected in increasing order.

MINIMUM NUMBER OF VARIABLES CARD:

This card is used if OPTION(1) is non-zero. It gives the minimum number of variables (including the dependent variable) at which stepwise deletion is to cease.

Its format is:

Columns	Format	Entry	Description
1 - 4	I4	K	The minimum number of variables (including the dependent variable on which the regression is to be performed. The stepwise deletion stops when M' is less than K. (K must be punched right justified.)

RESIDUAL START CARD:

This card is used if OPTION(2) is a two, and OPTION(1) is non-zero. It gives the number of variables (including the dependent variables) at which residual analysis is to be performed. Its format is the same as the MINIMUM NUMBER OF VARIABLES CARD.

OUTPUT:

Output to this program is a printed table of means, standard deviations, variance-covariance, the correlation matrix, multiple correlation coefficients (and degrees freedom), regression coefficients, regression weights, regression constants, standard errors of the regression weights and tests of significance, variables deleted (if any), residuals and sums of the squares of the residuals if requested, and the standard error of the estimate of the dependent variable.

Appendix: Sample Output

DCOSS REGRESSION

TITLE	INPUT ID	N	M	***** OPTIONS ****
FLOW (Q) VERSUS PH (63-69)	DATA	448	13	0 1 1 0 0 0 0 0 0 0

NUMBER OF SELECTED VARIABLES = 2

SELECTED VARIABLE INDICES
1 9

SELECTED VARIABLE IDENTIFIERS
PH FLOW

NUMBER OF SELECTED OBSERVATION PAIRS = 1

SELECTED OBSERVATION PAIRS
5- 23

```
DCOSS   REGRESSION

      TITLE                          INPUT ID        N     M    ***** OPTIONS *****
FLOW (Q) VERSUS PH     (68-83)       DATA          444    13    0 1 1 0 0 0 0 0 0 0

TOTAL NUMBER OF OBSERVATIONS =     19
NUMBER OF VALID OBSERVATIONS =     16

MEANS
   0.26470 01    0.21670 01
STANDARD DEVIATIONS
   0.92110 00    0.17130 01
```

```
00095 REGRESSION.....FLOW (Q) VERSUS PH    (68-69)

VARIANCE-COVARIANCE
ROW NUMBER      1
   0.84850 00  -0.55420 00
ROW NUMBER      2
  -0.55420 00   0.29350 01
```

DC095 REGRESSION.....FLOW (Q) VERSUS PH (68-69)

CORRELATION MATRIX
ROW NUMBER 1
0.10000 01 -0.35120 00
ROW NUMBER 2
-0.35120 00 0.10000 01

DCOSS REGRESSION.....FLOW (Q) VERSUS PH (AR-69)

VALUE OF DETRIMINANT = 0.2059E-83
MULTIPLE CORRELATION COEFFICIENT = 0.3512
TEST OF SIGNIFICANCE OF MULTIPLE CORRELATION COEFFICIENT = 0.14035E 01 DEGREES OF FREEDOM = 13
REGRESSION WEIGHTS
-0.3512E 00
REGRESSION COEFFICIENTS
-0.1689E 00
REGRESSION CONSTANT = 0.3056E 01
STANDARD ERROR OF REGRESSION WEIGHTS
2.5968E-01
TEST OF SIGNIFICANCE OF REGRESSION WEIGHTS
-1.3525E 00
SUM OF THE SQUARES OF THE RESIDUALS = 1.1157E 01
STANDARD ERROR OF THE ESTIMATE OF THE DEPENDENT VARIABLE = 9.2642E-01

Reproduced from best available copy.

00095 REGRESSION.....FLOW (J) VERSUS PH (SR=SS)

OBS	OBS TO	OBSERVED	COMPUTED	RESIDUAL	PER. ERROR	DATA TYPE NORMAL
5	CFS 2	2.0000E 00	1.9608E 00	3.9248E-02	1.9624E-02	
6	CFS 2	1.7000E 00	2.4707E 00	-7.7065E-01	-4.5332E-01	
7	CFS 2	1.9000E 00	2.6555E 00	-7.5550E-01	-3.9974E-01	
8	CFS 2	5.8000E 00	2.8672E 00	2.932BE 00	5.0565E-01	
10	CFS 2	2.5000E 00	2.4007E 00	1.9932E-01	7.9726E-02	
11	CFS 2	3.0000E 00	2.7728E 00	2.2719E-01	7.5729E-02	
12	CFS 2	2.7000E 00	2.9994E 00	-2.9944E-01	-1.1090E-01	
13	CFS 2	2.8000E 00	2.9428E 00	-1.4279E-01	-5.0994E-02	
14	CFS 2	2.8000E 00	3.0183E 00	-7.1832E-01	-3.1231E-01	
15	CFS 2	2.4000E 00	3.0872E 00	-6.8721E-01	-2.8550E-01	
16	CFS 2	2.2000E 00	2.5235E 00	-3.2353E-01	-1.4706E-01	
17	CFS 2	2.6000E 00	2.5840E 00	1.6037E-02	6.1681E-03	
18	CFS 2	2.4000E 00	2.4518E 00	-5.1766E-02	-2.1569E-02	
19	CFS 2	2.6000E 00	2.7917E 00	-1.9170E-01	-7.3731E-02	
20	CFS 2	2.3500E 00	2.1307E 00	2.1928E-01	9.3312E-02	
22	CFS 2	3.1000E 00	2.8389E 00	2.6109E-01	8.4221E-02	

MAXIMUM VALUE	MINIMUM VALUE	RANGE	MEAN ERROR	ABSOLUTE MEAN ERROR	ABSOLUTE MEAN PERCENTAGE ERROR
5.799999E 00	1.7000000E 00	4.099999E 00	7.748604E-07	4.868634E-01	1.687225E-01

INDEPENDENT VARIABLES LEFT IN THE EQUATION ARE -
9 (FLOW)

Reproduced from best available copy.